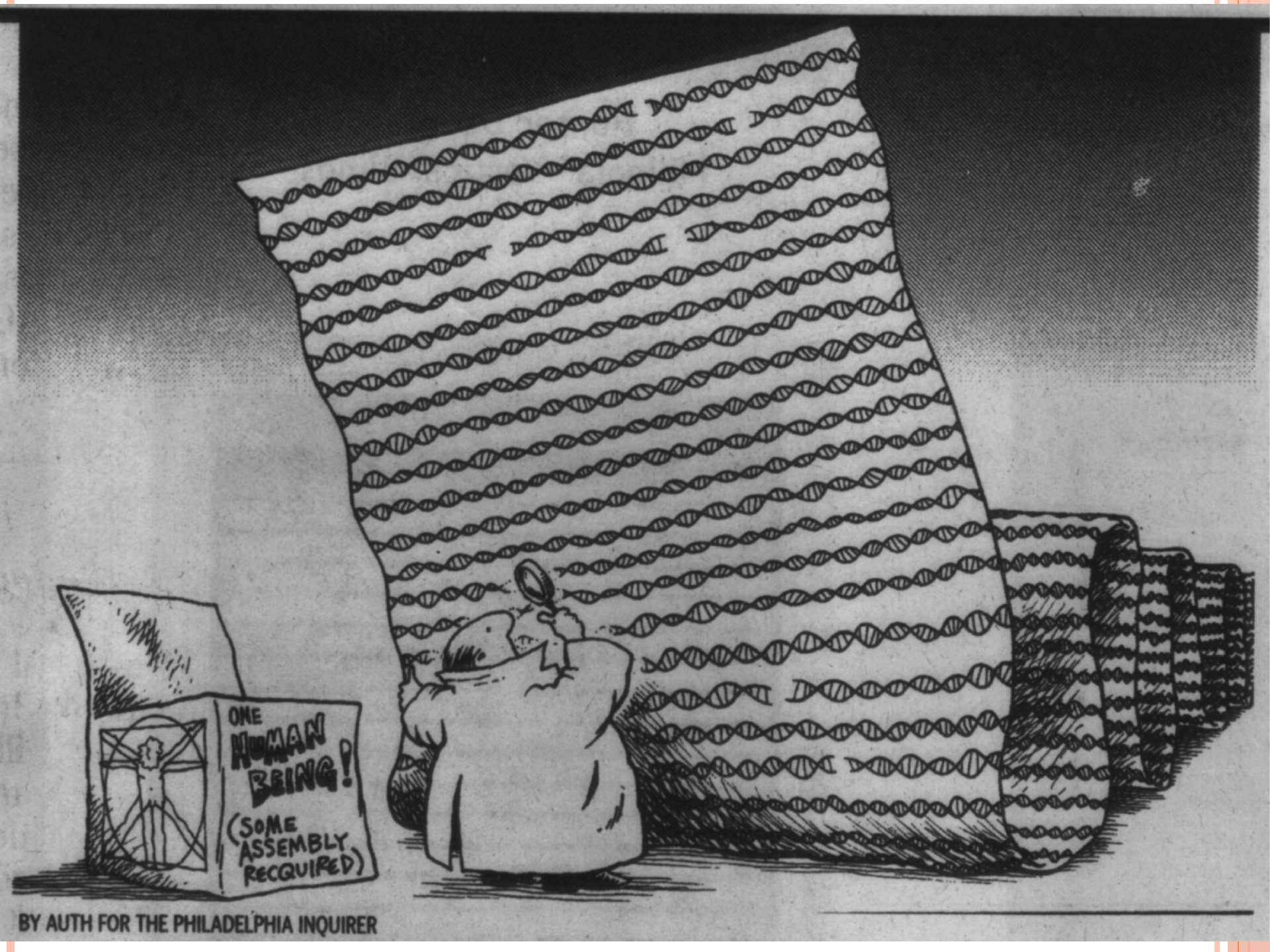
GENETICS AND MOLECULAR BIOLOGY FOR ENVIRONMENTAL ANALYSIS

MOLECULAR ECOLOGY

LESSON 14: ANALYSIS OF GENOMIC DATA

Prof. Alberto Pallavicini

pallavic@units.it


ONE
HUMAN
BEING!
(SOME
ASSEMBLY
REQUIRED)
BY AUTH FOR THE PHILADELPHIA INQUIRER

VOCABULARY

- Fragment library: a short insert (270bp) library with overlapping ends. Aka std library
- Long insert library: A 4-8kb library where only 100 bp on each end are sequenced. Aka mate pair library
- Contig: A contiguous sequence of DNA
- Scaffold: One or more contigs linked together by unknown sequence
- Captured gap: A gap within a scaffold. The order and orientation of the contigs spanning the gap is known

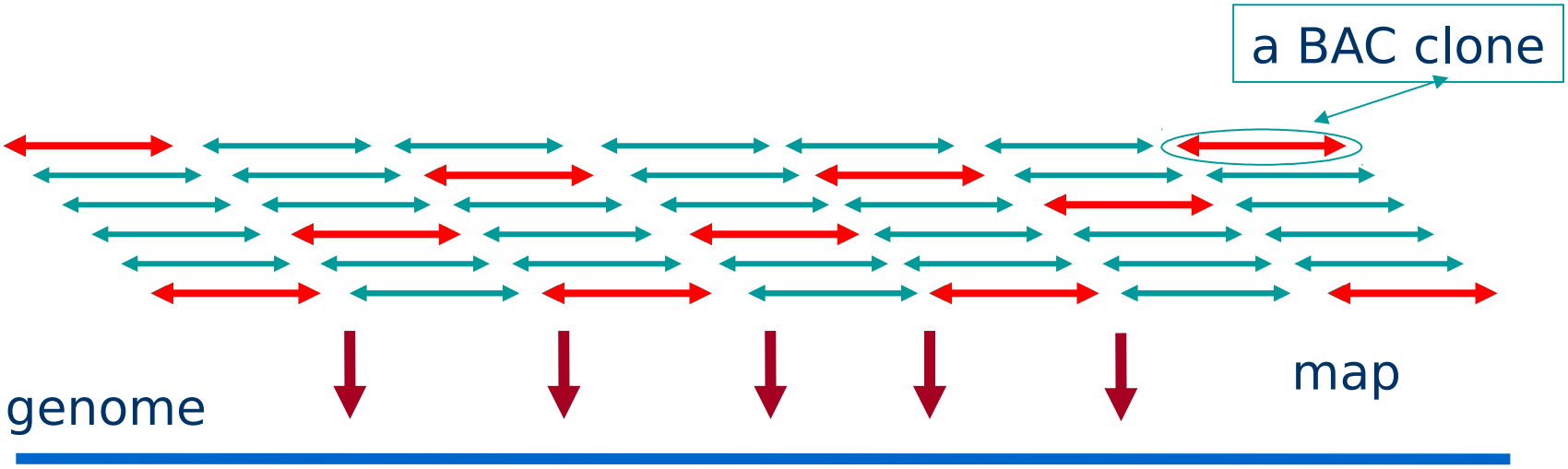
STORY OF STRATEGIES FOR WHOLE-GENOME SEQUENCING

- Hierarchical – Clone-by-clone yeast, worm, human
  - Break genome into many long fragments
  - Map each long fragment onto the genome
  - Sequence each fragment with shotgun
- Online version of hierarchical – Walking rice genome
  i. Break genome into many long fragments
  ii. Start sequencing each fragment with shotgun
  iii. Construct map as you go
- Whole Genome Shotgun fly, human, mouse, rat, fugu

One large shotgun pass on the whole genome

HIERARCHICAL SEQUENCING

HIERARCHICAL SEQUENCING STRATEGY



1. Obtain a large collection of BAC clones
2. Map them onto the genome (Physical Mapping)
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

ONLINE CLONE-BY-CLONE
THE WALKING METHOD

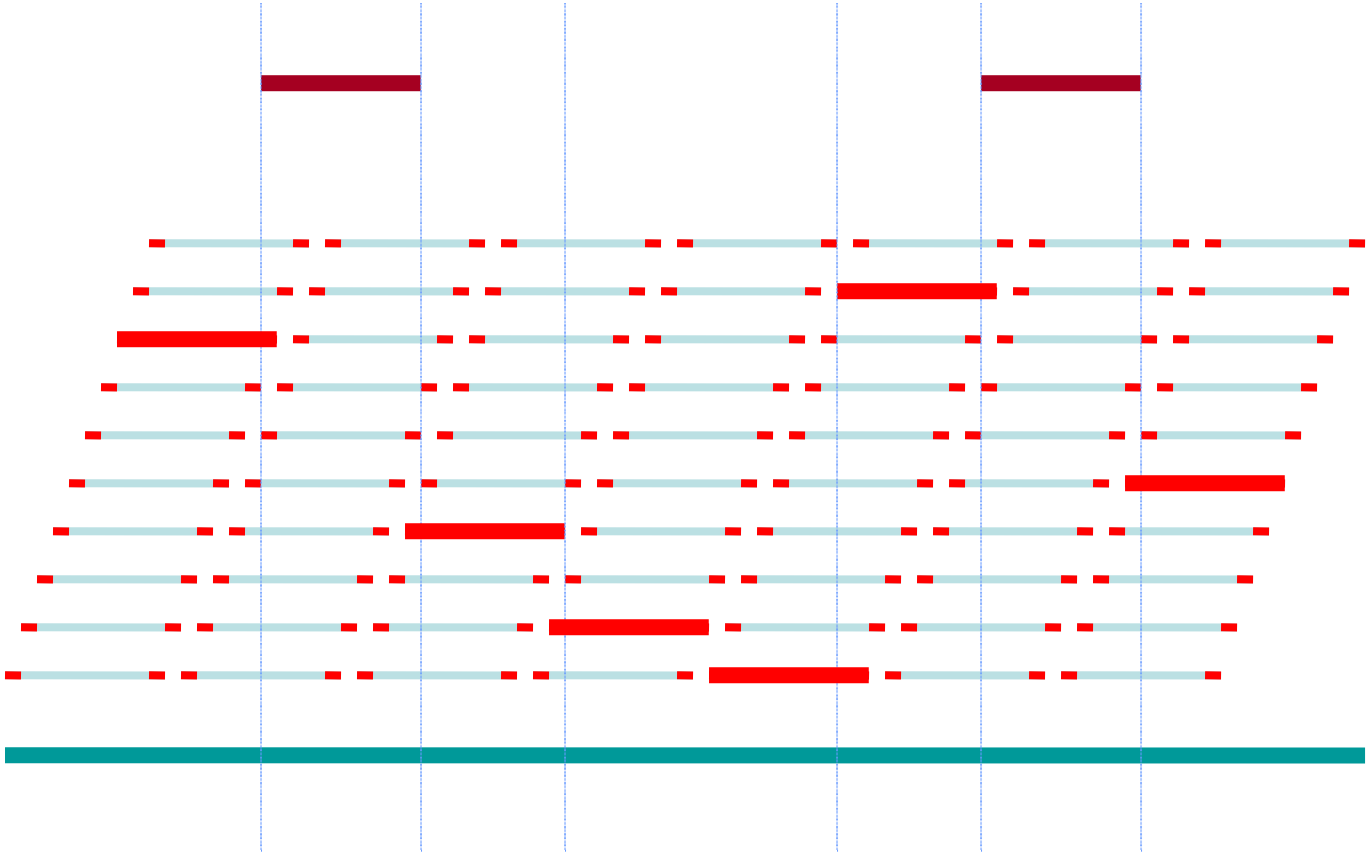
THE WALKING METHOD

1. Build a very redundant library of BACs with sequenced clone-ends

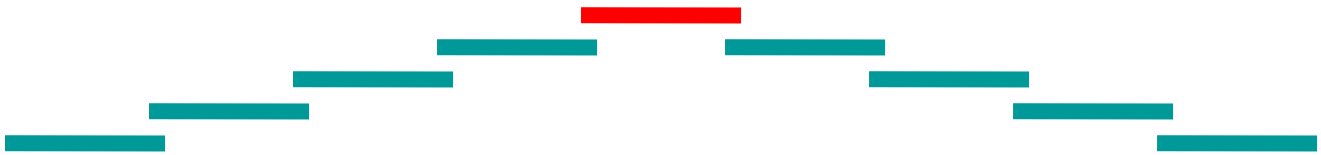
2. Sequence some “seed” clones

3. “Walk” from seeds using clone-ends to pick library clones that extend left & right

Walking: An Example



WALKING OFF A SINGLE SEED



- Low redundant sequencing
- Many sequential steps

Walking off a single clone is impractical

Cycle time to process one clone: 1-2 months

1. Grow clone
2. Prepare & Shear DNA
3. Prepare shotgun library & perform shotgun
4. Assemble in a computer
5. Close remaining gaps

A mammalian genome would need **15,000** walking steps !

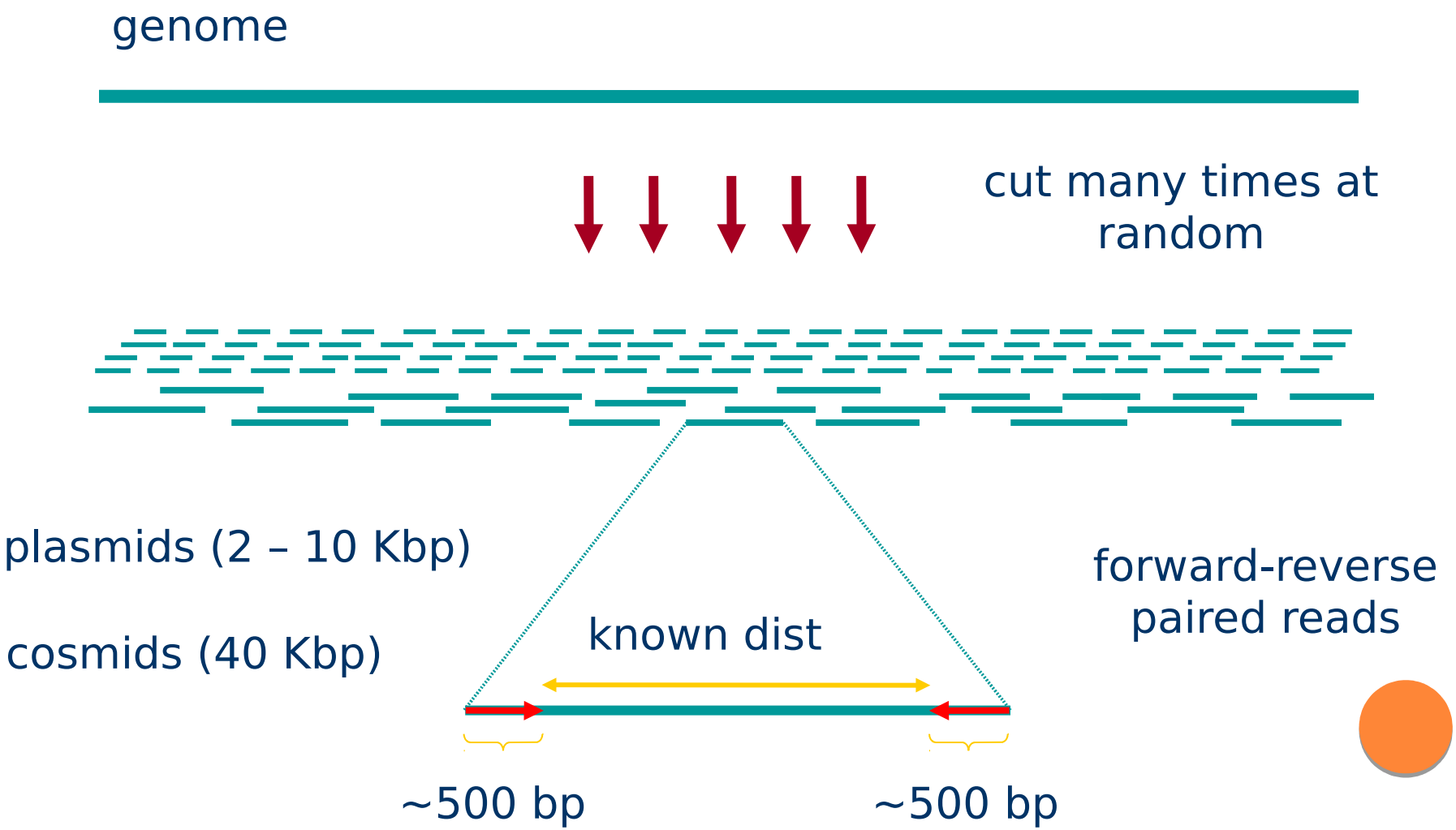
WALKING OFF SEVERAL SEEDS IN PARALLEL

- Few sequential steps
- Additional redundant sequencing

In general, can sequence a genome in ~5 walking steps, with <20% redundant sequencing

WHOLE-GENOME SHOTGUN SEQUENCING

WHOLE GENOME SHOTGUN SEQUENCING

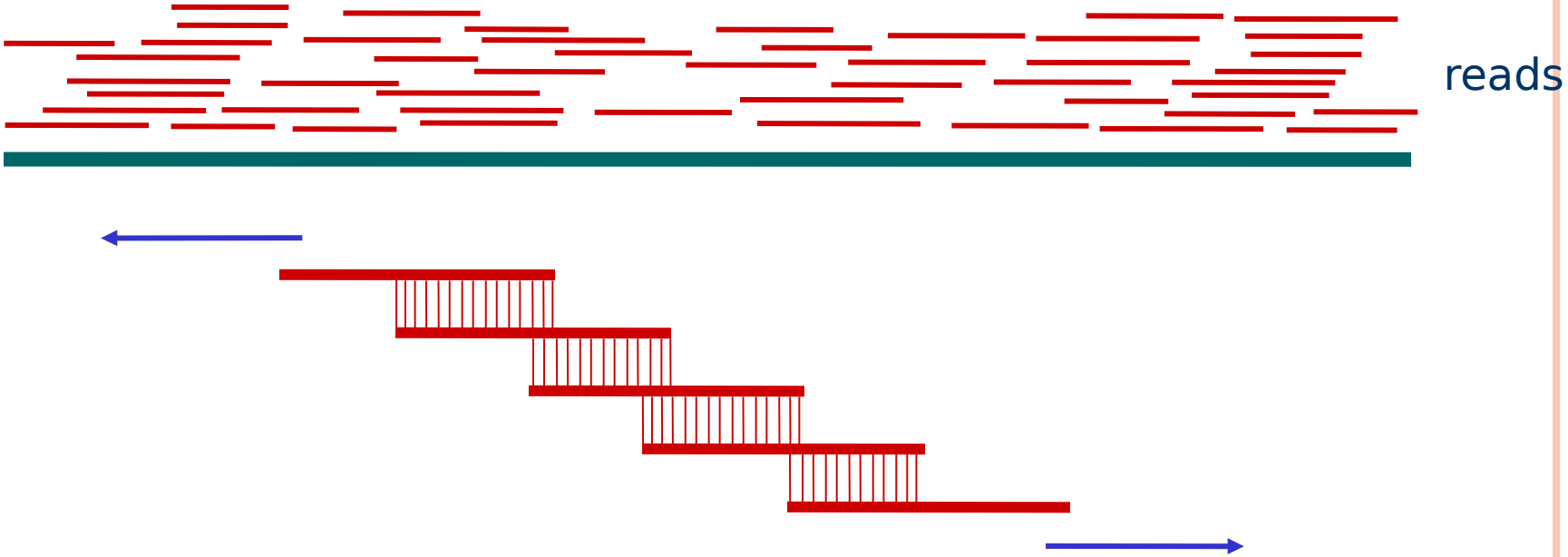


HIERARCHICAL SEQUENCING VS. WHOLE GENOME SHOTGUN

- Hierarchical Sequencing
  - Advantages: Easy assembly
  - Disadvantages:
    - Build library & physical map;
    - Redundant sequencing
- Whole Genome Shotgun (WGS)
  - Advantages: No mapping, no redundant sequencing
  - Disadvantages: Difficult to assemble and resolve repeats

Whole Genome Shotgun appears to get more popular…
Mainly now with ngs!

FRAGMENT ASSEMBLY



Cover region with ~7-fold redundancy

Overlap reads and extend to reconstruct the original genomic region

READ COVERAGE

Length of genomic segment: G

Number of reads: N

Length of each read: L

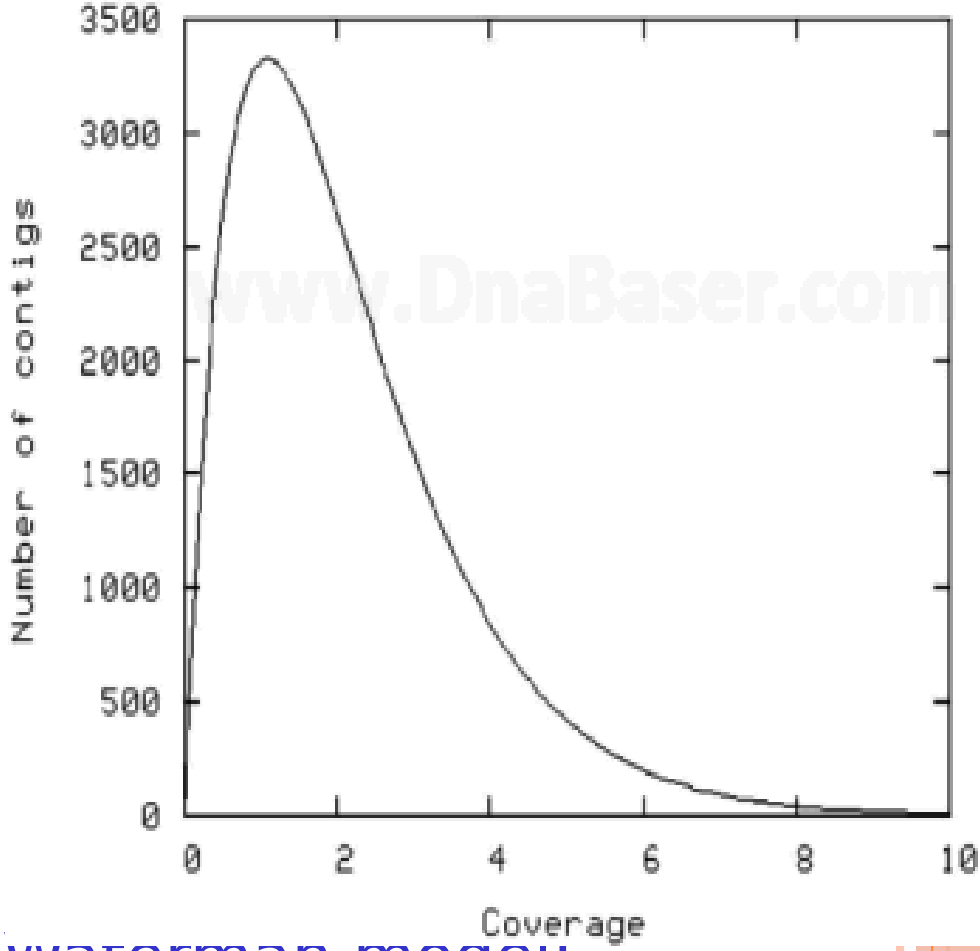
<u>Definition:</u> Coverage $C = NL/G$

N50

An N50 contig size of N means that 50% of the assembled bases are contained in contigs of length N or larger.

N50 sizes are often used as a measure of assembly quality because they capture how much of the genome is covered by relatively large contigs.

ENOUGH COVERAGE



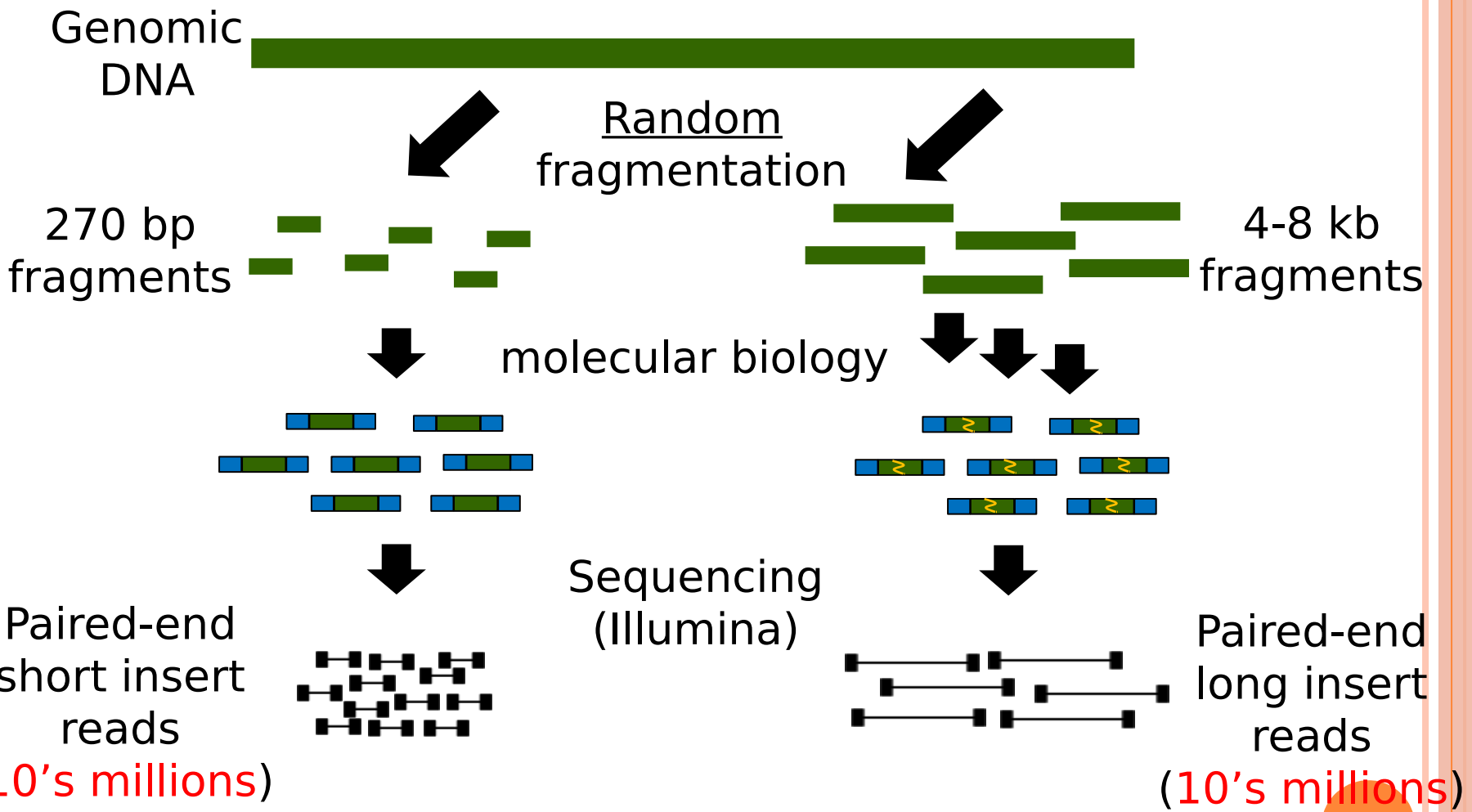
How much coverage is enough?

According to the Lander-Waterman model:

Assuming uniform distribution of reads, C=7 results in 1 gap per 1,000 nucleotides (remember this was corrrect for Sanger reads)

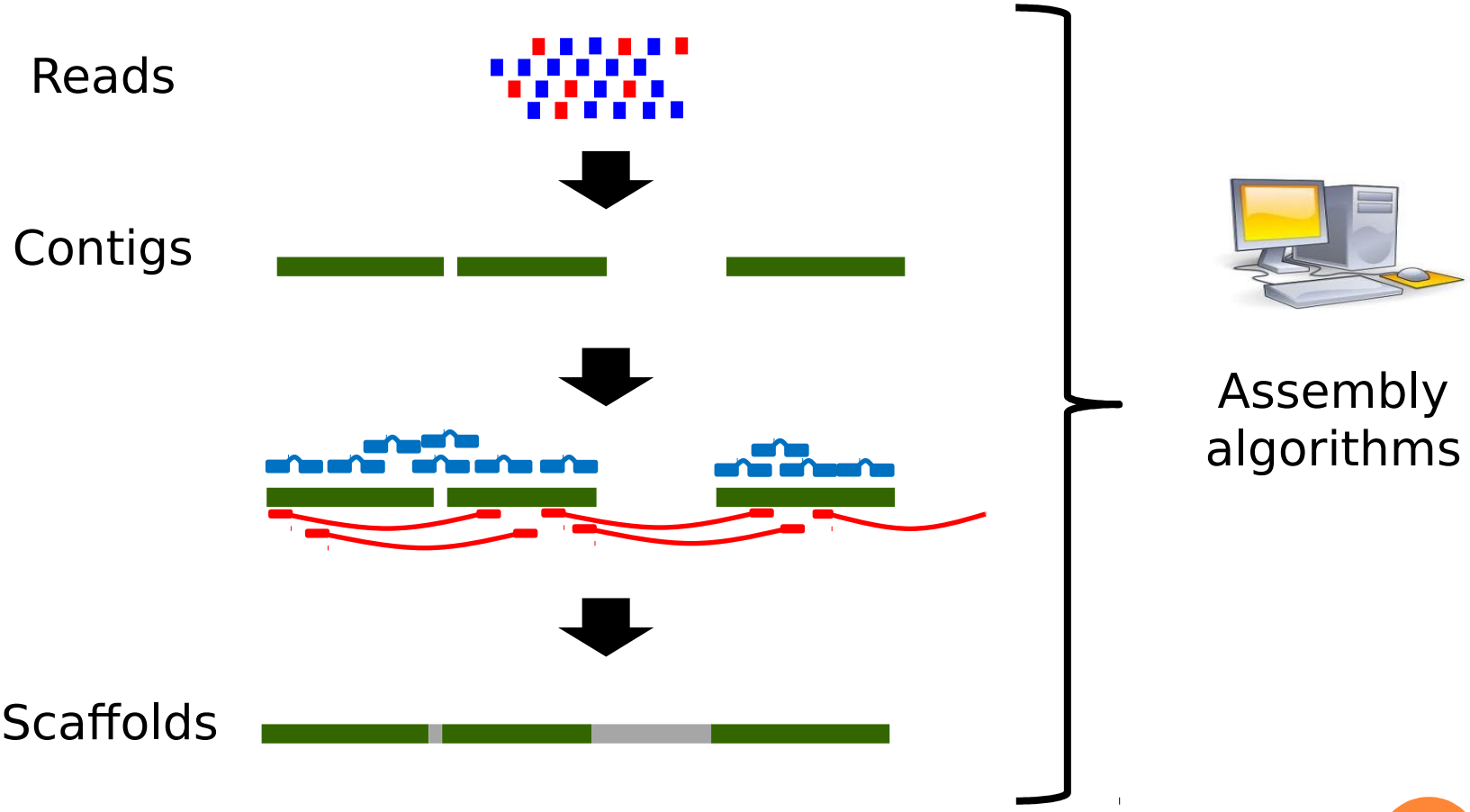
2. Introduction to short-read genome sequencing and assembly
   - **Short read sequencing and assembly basics**
   - Short read assembly - De Bruijn graph example
   - Short read assembly – Scaffolding

SHORT READ GENOME SEQUENCING



How do we assemble this data back into a genome?

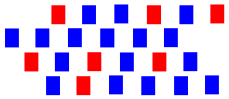
ASSEMBLY OUTLINE



https://omictools.com/genome-assembly-category

ASSEMBLY OUTLINE

Reads



‘De Bruijn’ assembly

Contigs

Scaffolds

2. Introduction to short-read genome sequencing and assembly
   - Short read sequencing and assembly basics
   - **Short read assembly - De Bruijn graph example**
   - Short read assembly – Scaffolding

DE BRUIJN EXAMPLE

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity,.... “

Dickens, Charles. A Tale of Two Cities. 1859. London: Chapman Hall

Era il tempo migliore e il tempo peggiore, la stagione della saggezza e la stagione della follia, l'epoca della fede e l'epoca dell'incredulità, il periodo della luce e il periodo delle tenebre, la primavera della speranza e l'inverno della disperazione.

Example courtesy of J. Leipzig 2010

DE BRUIJN EXAMPLE

itwasthebestoftimesitwastheworstoftimesitwastheageofwisdomitwastheageoffoolishness..

Generate random ‘reads’

How do we assemble?

fincreduli geoffoolis Itwasthebe Itwasthebe geofwisdom itwastheep epochofinc timesitwas stheepocho nessitwast wastheageo theepochof stheepocho hofincredu estoftimes eoffoolish lishnessit hofbeliefi pochofincr itwasthewo twastheage toftimesit domitwasth ochofbelie eepochofbe eepochofbe astheworst chofincred theageofwi iefitwasth ssitwasthe astheepoch efitwasthe wisdomitwa ageoffooli twasthewor ochofbelie sdomitwast sitwasthea eepochofbe ffoolishne eofwisdomi hebestofti stheageoff twastheepo eworstofti stoftimesi theepochof esitwasthe heepochofi theepochof sdomitwast astheworst rstoftimes worstoftim stheepocho geoffoolis ffoolishne timesitwas lishnessit stheageoff eworstofti orstoftime fwisdomitw wastheageo heageofwis incredulit ishnessitw twastheepo wasthewors astheepoch heworstoft ofbeliefit wastheageo heepochofi pochofincr heageofwis stheageofw fincreduli astheageof wisdomitwa wastheageo astheepoch olishnessi astheepoch itwastheep twastheage wisdomitwa fbeliefitw bestoftime epochofbel theepochof sthebestof lishnessit hofbeliefi Itwasthebe ishnessitw sitwasthew ageofwisdo twastheage esitwasthe twastheage shnessitwa fincreduli fbeliefitw theepochof mesitwasth domitwasth ochofbelie heageofwis oftimesitw stheepocho bestoftime twastheage foolishnes ftimesitwa thebestoft itwastheag theepochof itwasthewo ofbeliefit bestoftime mitwasthea imesitwast timesitwas orstoftime estoftimes twasthebes stoftimesi sdomitwast wisdomitwa theworstof astheworst sitwasthew theageoffo eepochofbe

...etc. to 10’s of millions of reads

Traditional all-vs-all assemblers fail due to immense computational resources (scales with number of reads^2)

A million (10^6) reads requires a trillion (10^{12}) pairwise alignments

De Bruijn solution:

Represent the data as a graph (scales with genome size)

DE BRUIJN EXAMPLE

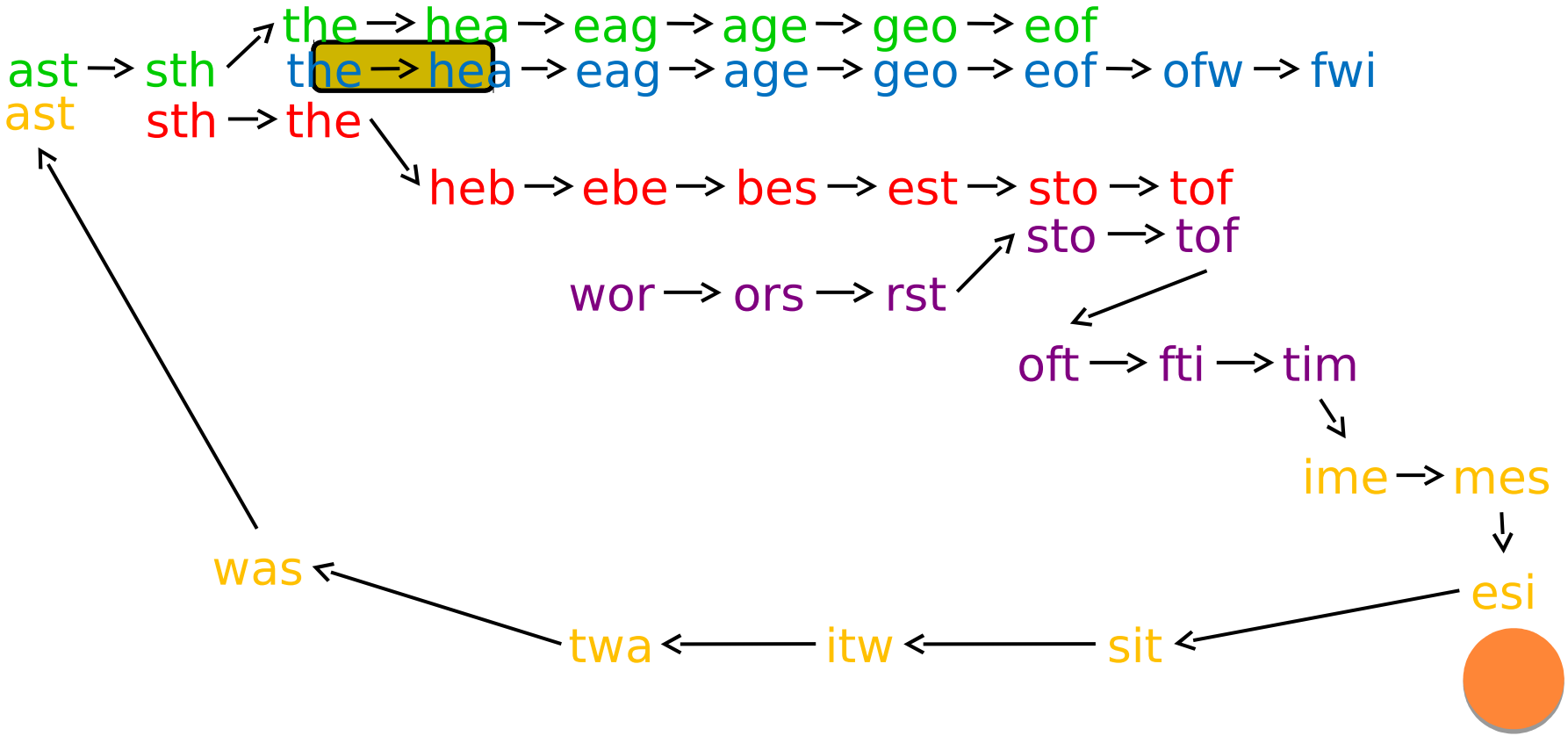
Step 1:
Convert reads into “Kmers”

Kmer: a substring of defined length

.....etc for all reads in the dataset

DE BRUIJN EXAMPLE

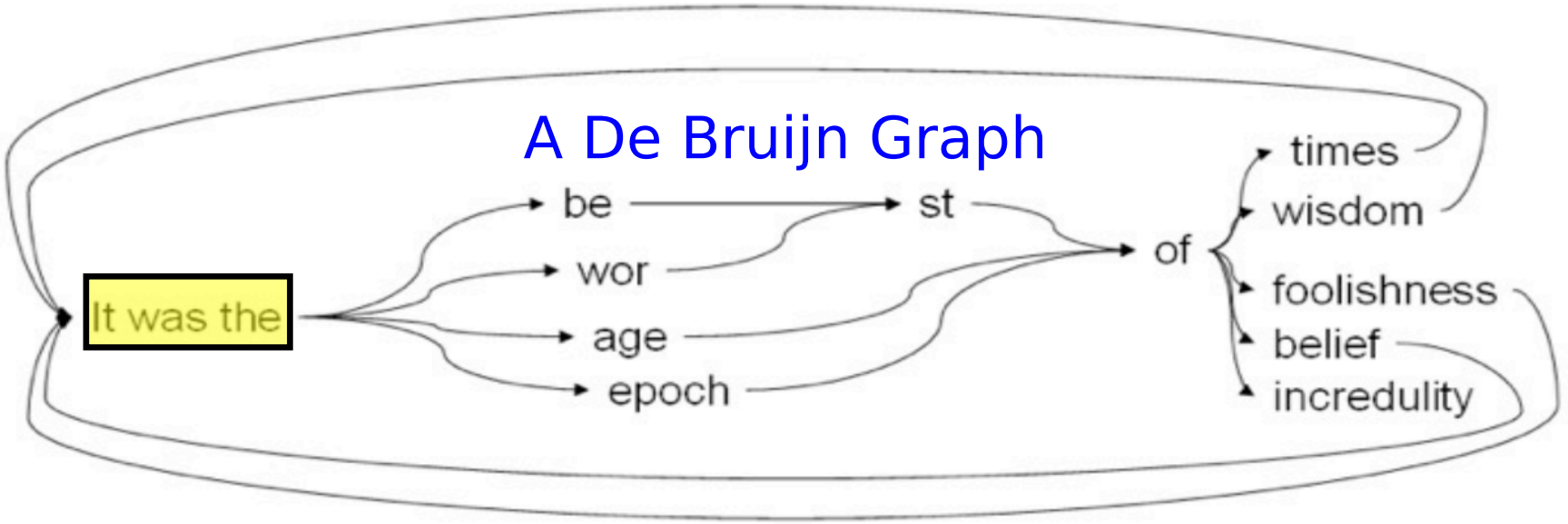
Step 2:
Build a De-Bruijn graph from the kmers



…..etc for all ‘kmers’ in the dataset

DE BRUIJN EXAMPLE

Step 3:
Simplify the graph as much as possible:

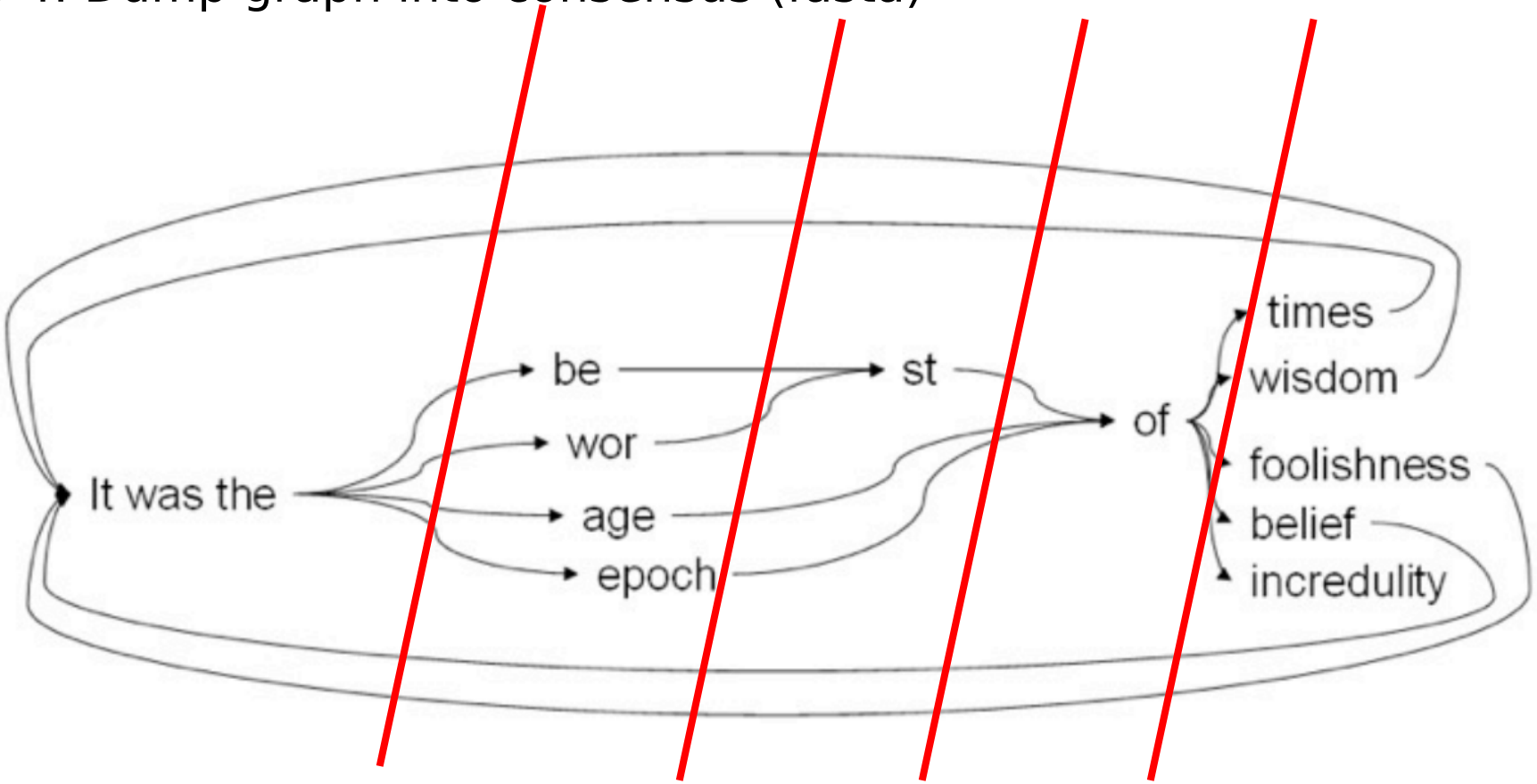


De Bruijn assemblies ‘broken’ by repeats longer than kmer

“It was the best of times, it was the worst of times, It was the age of wisdom, It was the age of foolishness, It was the epoch of belief, it was the epoch of incredulity,.... “

DRAWBACK OF DE BRUIJN APPROACH

Step 4: Dump graph into consensus (fasta)



No single solution!

Break graph to produce final assembly

KMER SIZE IS AN IMPORTANT PARAMETER IN DE BRUIJN ASSEMBLY

The final assembly (k=3)

wor times itwasthe foolishness st wisdom

incredulity age epoch be of belief

Repeat with a longer “kmer” length

A better assembly (k=20)

itwasthebestoftimesitwastheworstoftimesitwastheageofwisdomitwastheageoffoolis…

Why not always use longest ‘k’ possible?

Higher k-mer sizes

Having larger sized k-mers will increase the amount of edges in the graph, which in turn, will increase the amount of memory needed to store the DNA sequence.

Larger k-mers also run a higher risk of not having outward vertices from every k-mer. This is due to larger k-mers increasing the risk that it will not overlap with another k-mer by $k - 1$.

BUBBLE RESOLUTION

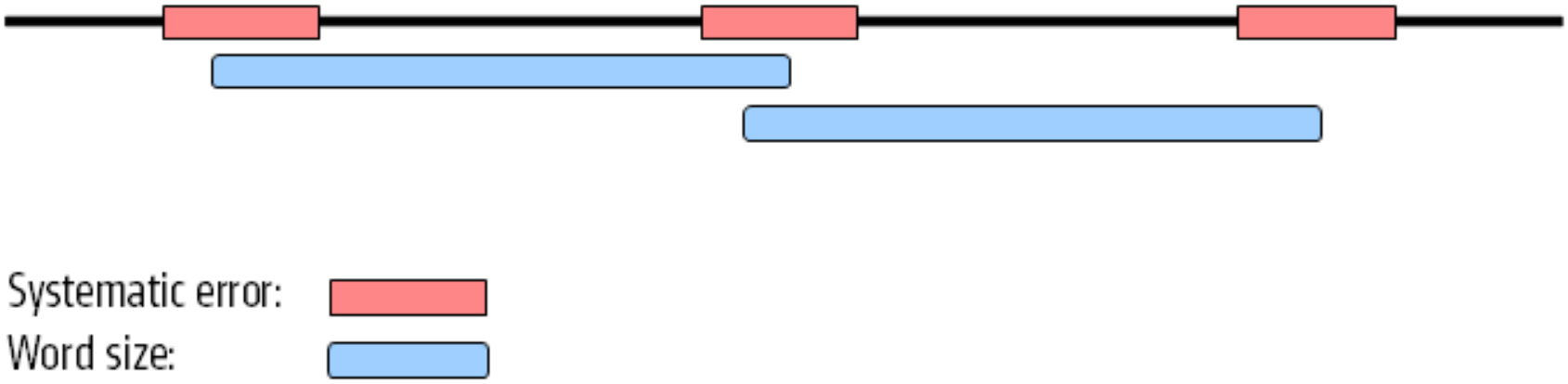
- Before the graph structure is converted to contig sequences, bubbles are resolved. As mentioned previously, a bubble is defined as a bifurcation in the graph where a path furcates into two nodes and then merge back into one.

- In this simple case the assembler will collapse the bubble and use the route through the graph that has the highest coverage of reads.

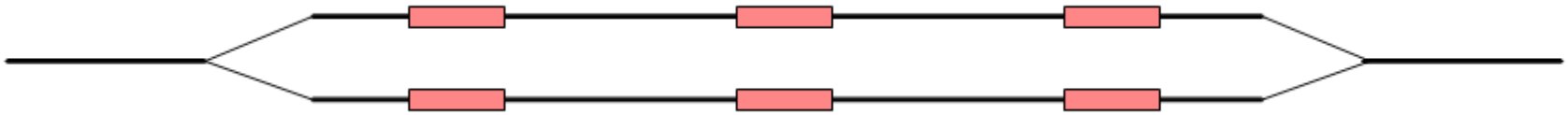
- This figure shows an example of a data set where the reads have systematic errors. Some reads include five As and others have six. This is a typical example of the homopolymer errors.

SYSTEMATIC ERRORS

- When these reads are assembled, this site will give rise to a bubble in the graph. This is not a problem in itself, but if there are several of these sites close together, the two paths in the graph will not be able to merge between each site. This happens when the distance between the sites is smaller than the word size used.



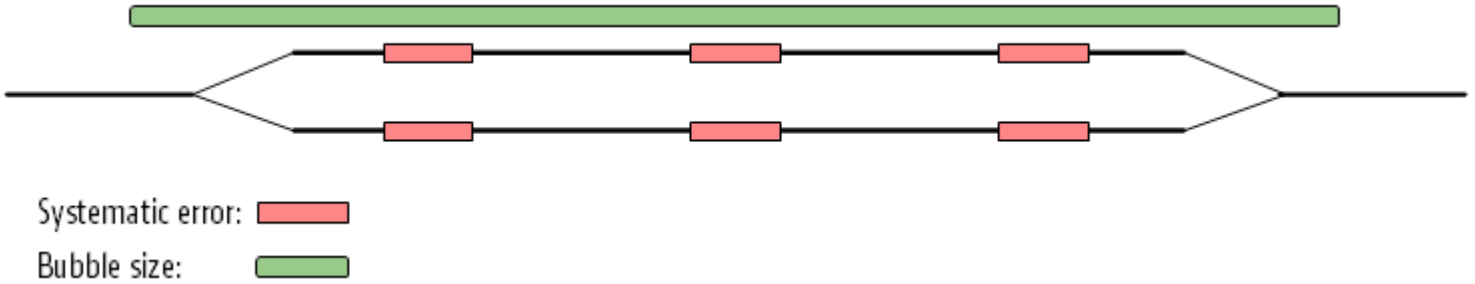
- In this case, the bubble will be very large because there are no complete words in the regions between the homopolymer sites, and the graph will look like:



- If the bubble is too large, the assembler will have to break it into several separate contigs instead of producing one single contig.

MAXIMUM BUBBLE SIZE

- The maximum size of bubbles that the assembler should try to resolve can be set by the user.



- Please keep in mind that increasing the bubble size also increases the change of misassemblies.

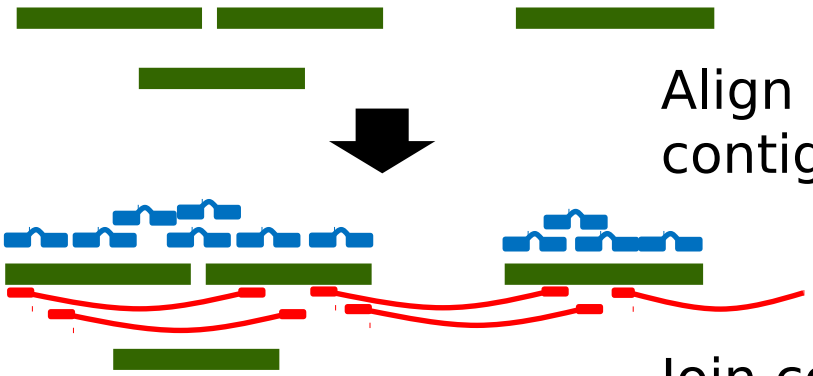
2. Introduction to short-read genome sequencing and assembly
   - Short read sequencing and assembly basics
   - Short read assembly - De Bruijn graph example
   - **Short read assembly – Scaffolding**

SCAFFOLDING

Reads

‘De Bruijn’ assembly

Contigs



Align reads to DeBruijn contigs

Join contigs using evidence from paired end data

Scaffolds

(An assembly)

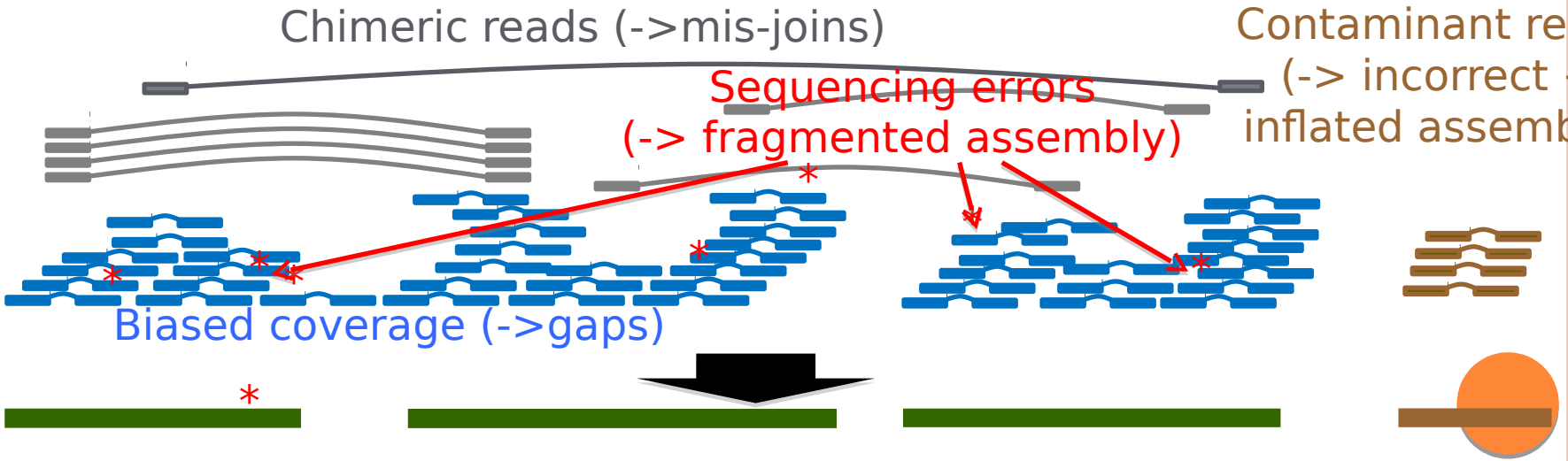
“Captured” gaps caused by repeats. Represented by “NNN” in assembly

REAL LIFE ASSEMBLY IS MESSY!

Assembly in theory

Uniform coverage, no errors, no contamination

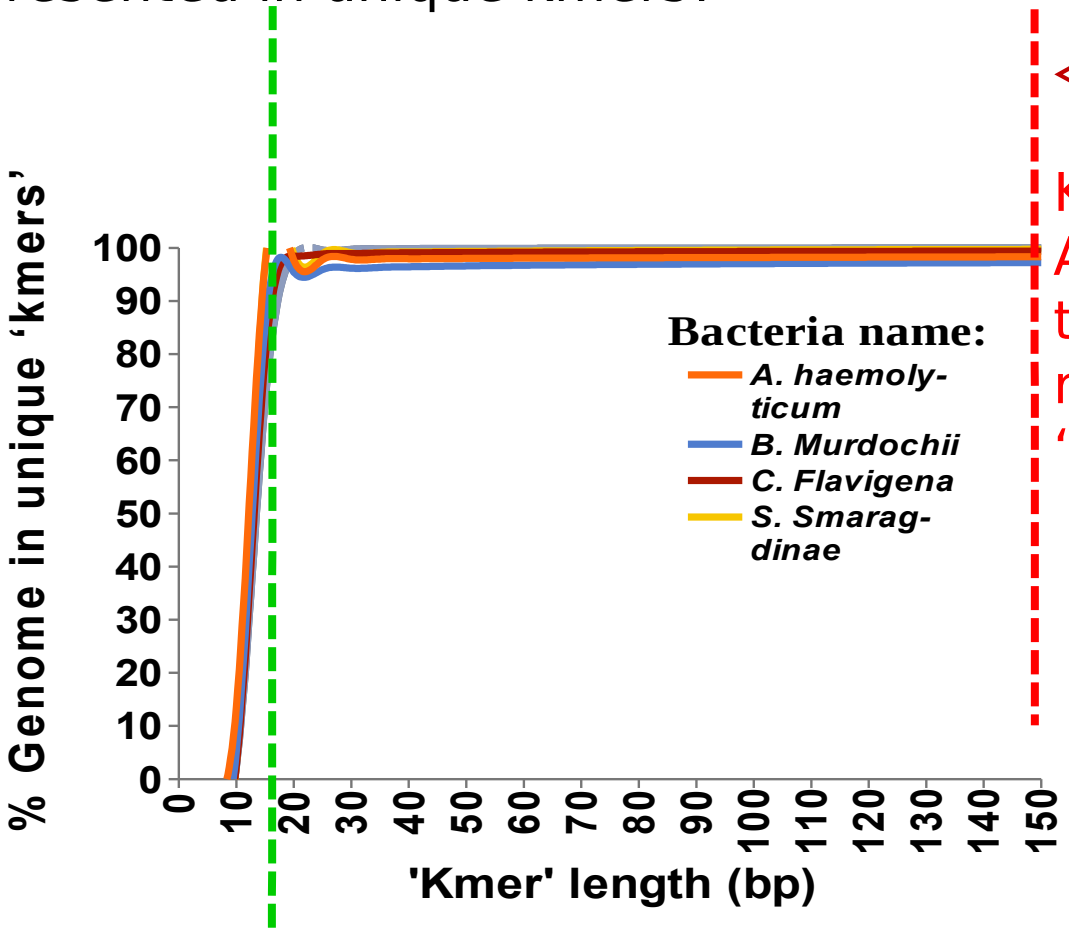
Assembly in reality



Worse than predicted assemblies!

SIMULATED DE BRUIJN ASSEMBLY FOR SIX 'KNOWN'MICROBIAL GENOMES

What fraction of a genome should we be able to assemble, that is, can be represented in unique kmers?



← 97 %

Kmer = 150

A small fraction of the genome remains 'unassemblable'

Kmer =30, most of the genome CAN be assembled

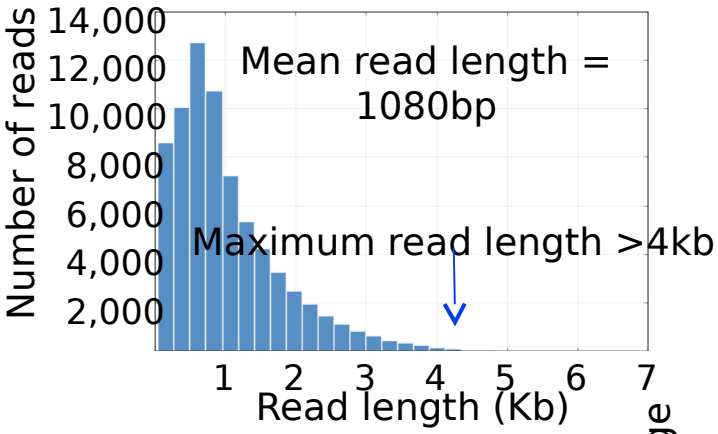
Matt Blow

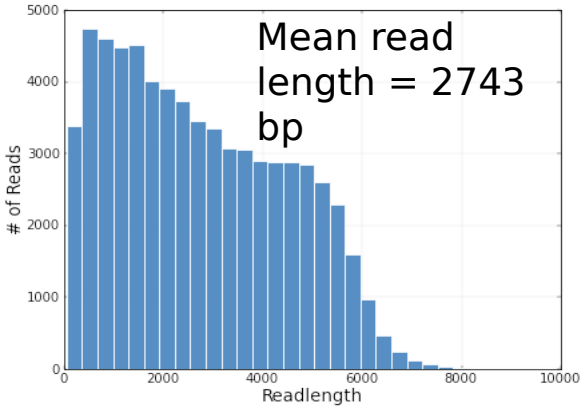
1. Vocabulary introduction
2. Introduction to short-read genome sequencing and assembly
3. Practical experience of short read genome assembly
4. Improving genome assembly using 3rd generation sequencing

PACIFIC BIOSCIENCES SEQUENCER

Long reads from "3rd generation" Pacific Biosciences sequencer hold promise for improving short-read based assemblies

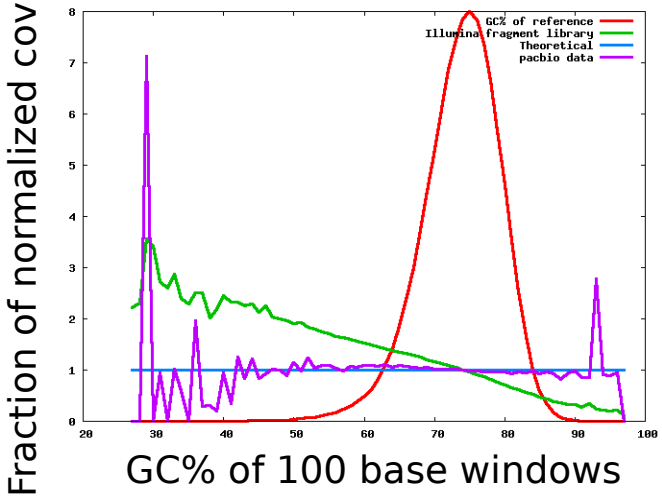
"Early" data





Low coverage bias

Reduced sensitivity to G+C rich regions compared to illumina chemistry



High error rate

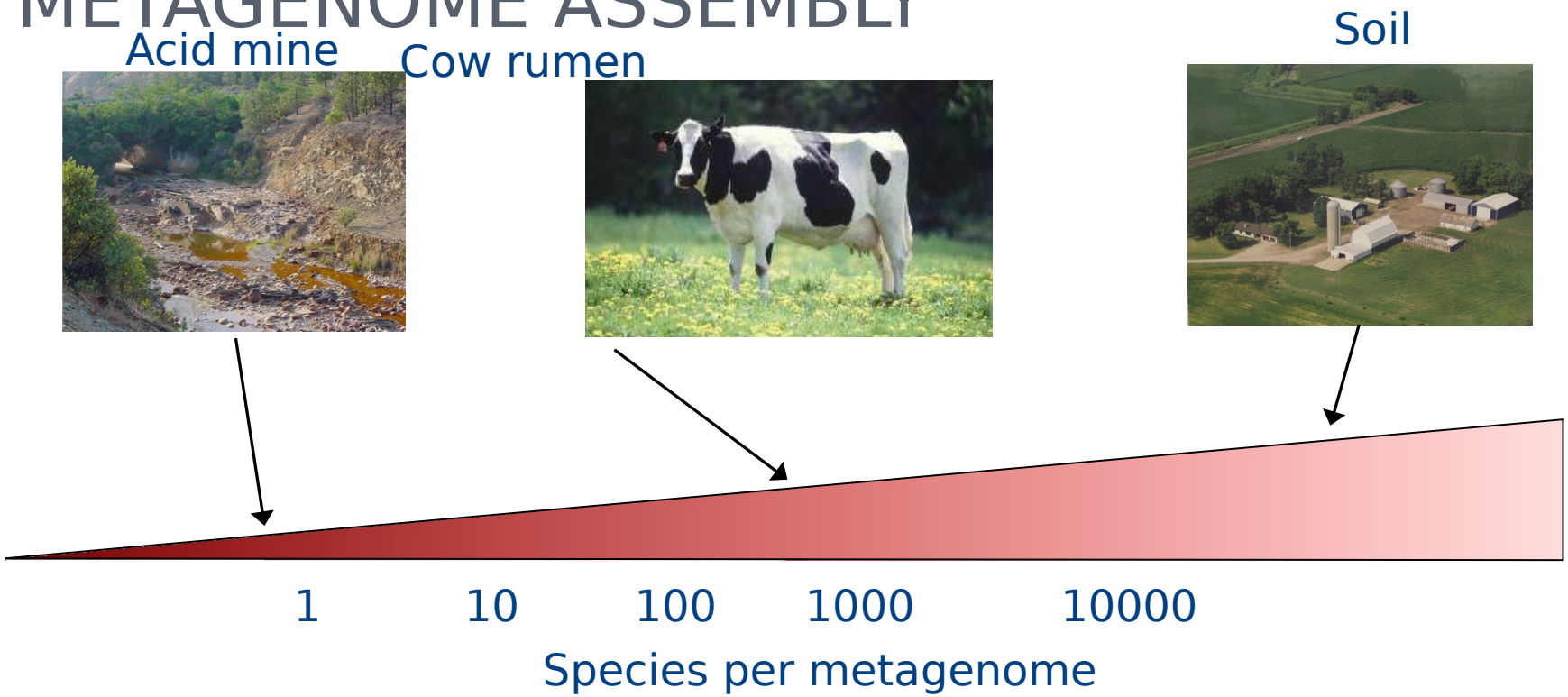
Up to 15% error rate

In-del errors

SUMMARY

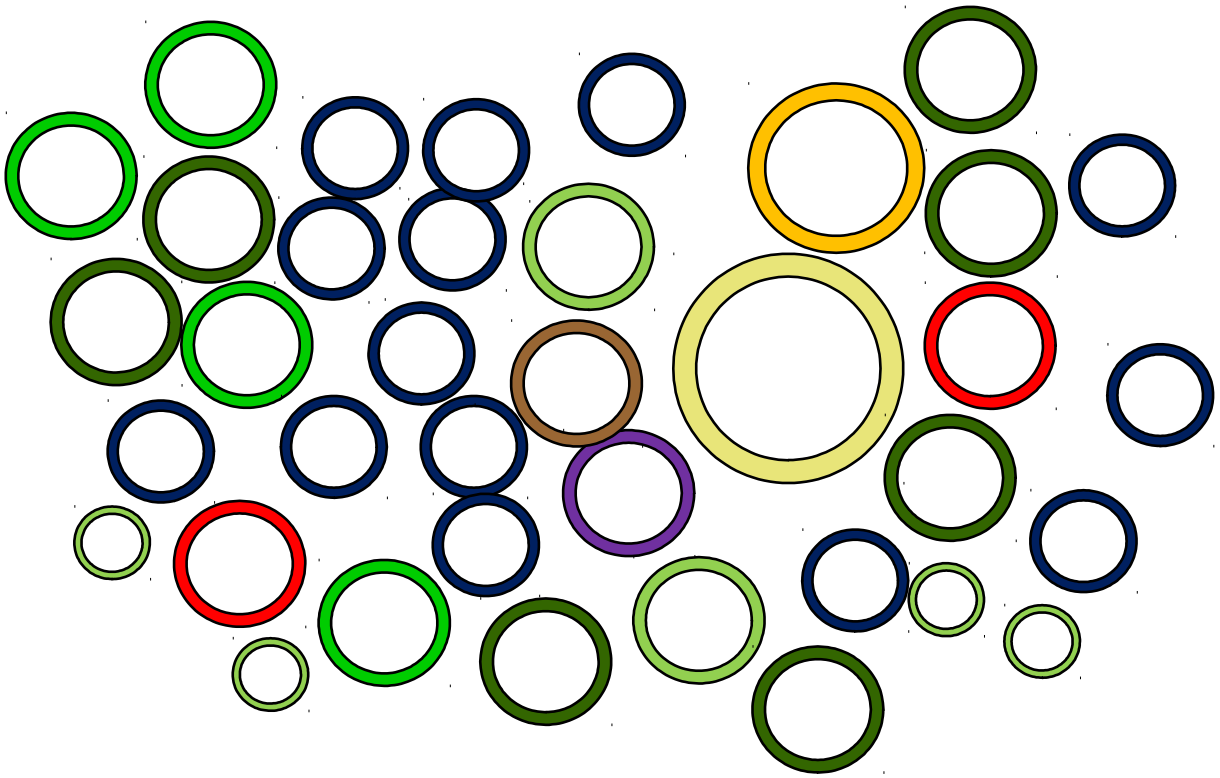
- High quality genome sequencing using only short-reads is within reach
- Short-read microbial genomes assemblies are minimally fragmented and contain the vast majority known genes
- Third-generation sequencing may provide an inexpensive path to finished genomes

METAGENOME ASSEMBLY



-All challenges of isolate genome assembly remain
-Extra challenges from diversity and different abundance of constituent genomes
- The same strategies as isolate assembly can be used, but many heuristics fail for metagenomes

METAGENOME ASSEMBLY IS AN ONGOING CHALLENGE



-All challenges of isolate genome assembly remain
-Extra challenges from diversity and different abundance of constituent genomes
- The same strategies as isolate assembly can be used, but many heuristics fail for metagenomes

USEFUL REVIEWS

- Miller JR, Koren S, Sutton G. , Assembly algorithms for next-generation sequencing data. Genomics. 2010 Jun;95(6):315-27.
- Mihai Pop, Genome assembly reborn: recent computational challenges. Brief Bioinform (2009) 10 (4):354-366.

The sequence and *de novo* assembly of the giant panda genome

nature
International weekly journal of science

Article

Nature **463**, 311-317 (21 January 2010) | doi:10.1038/nature08696; Received 19 August 2009; Accepted 24 November 2009; Published online 13 December 2009

There is a Corrigendum (25 February 2010) associated with this document.

The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li[1,2,27], Wei Fan[1,27], Geng Tian[1,3,27], Hongmei Zhu[1,27], Lin He[4,5,27], Jing Cai[3,6,27], Quanfei Huang[1], Qingle Cai[1,7], Bo Li[1], Yinqi Bai[1], Zhihe Zhang[8], Yaping Zhang[6], Wen Wang[6], Jun Li[1], Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1], Zhaolei Zhang[11], Rasmus Nielsen[12], Dawei Li[1], Wanjun Gu[13], Zhentao Yang[1], Zhaoling Xuan[1], Oliver A. Ryder[14], Frederick Chi-Ching Leung[15], Yan Zhou[1], Jianjun Cao[1], Xiao Sun[16], Yonggui Fu[17], Xiaodong Fang[1], Xiaosen Guo[1], Bo Wang[1], Rong Hou[8], Fujun Shen[8], Bo Mu[1], Peixiang Ni[1], Runmao Lin[1], Wubin Qian[1], Guodong Wang[3,6], Chang Yu[1], Wenhui Nie[6], Jinhuan Wang[6], Zhigang Wu[1], Huiqing Liang[1], Jiumeng Min[1,7], Qi Wu[9], Shifeng Cheng[1,7], Jue Ruan[1,3], Mingwei Wang[1], Zhongbin Shi[1], Ming Wen[1], Binghang Liu[1], Xiaoli Ren[1], Huisong Zheng[1], Dong Dong[11], Kathleen Cook[11], Gao Shan[1],

FULL TEXT
Previous | Next
Table of contents
Download PDF
Send to a friend
CrossRef lists 25 articl this article
Scopus lists 37 article

Panda genome completed using short-read sequencing

By John Timmer | Last updated December 14, 2009 5:15 PM

Over the weekend, *Nature* released a paper that describes the genome of the giant panda. The results do tell us something about the animal itself—it's a vegetarian that looks like a carnivore—but is probably more notable for how it was obtained. It's the first genome to be completed using sequence from the relatively new Illumina platform, which typically reads less than 100 bases from each piece of DNA, but can read massive amounts of DNA in parallel.

Illumina isn't one of the technologies I've covered yet in our DNA sequencing series, but it's similar to some of the other ones, like Complete Genomics, in that the read length is very short (an average of 52 bases for the panda). But Illumina machines produce staggering amounts of these short reads in each run, and provided the authors with enough sequence to read each base, on average, 73 times. Figuring out how all these tiny pieces (all 3.3 billion of them) overlap is a serious computational problem, but the authors used a software package called SOAP that's specifically designed to handle these short reads.

The genome that resulted looks a lot like the one from the only other carnivore sequenced, the dog. Despite

37 paired-end sequence libraries, read length=52bp on average, average depth coverage per base =73